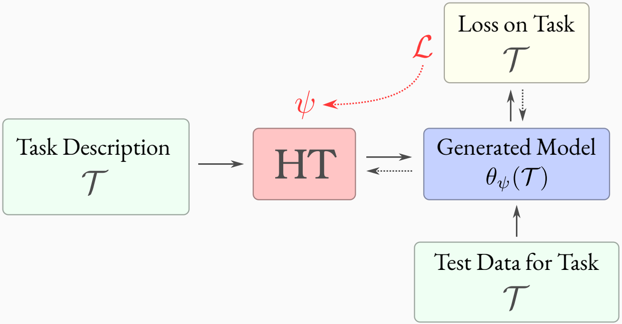
HyperTransformer: Model Generation for Supervised and Semi-Supervised Few-Shot Learning

Andrey Zhmoginov, Mark Sandler, Max Vladymyrov (Google Research)

ICML 2022

Core Idea: Learning to Generate Weights

- **Core idea**:

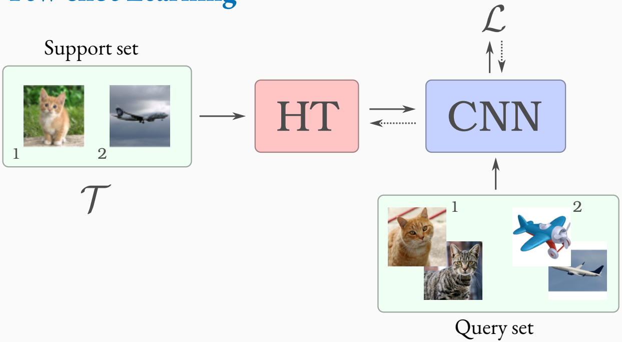


- Finding $\theta_\psi(T)$ solving

$$\underset{\psi}{\operatorname{argmin}}\, \mathbb{E}_T \left[\mathcal{L}_T(\theta_\psi(T))\right]$$

- **Related Work**:
  - D. Ha, A. Dai, Q. V. Le, *HyperNetworks*
  - B. Knyazev, et al., *Parameter Prediction for Unseen Deep Architectures*

- **Application to Few-shot Learning**

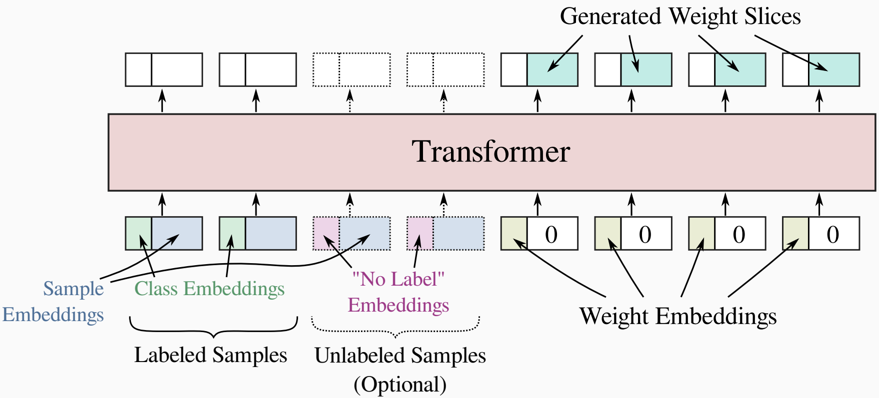


- Decoupling the complexities of the *model generator* (HT) and the *generated model*
  - Run generated models efficiently *if tasks don't change often*
- Versatility of "task descriptions" $\longrightarrow$ supervised and semi-supervised learning

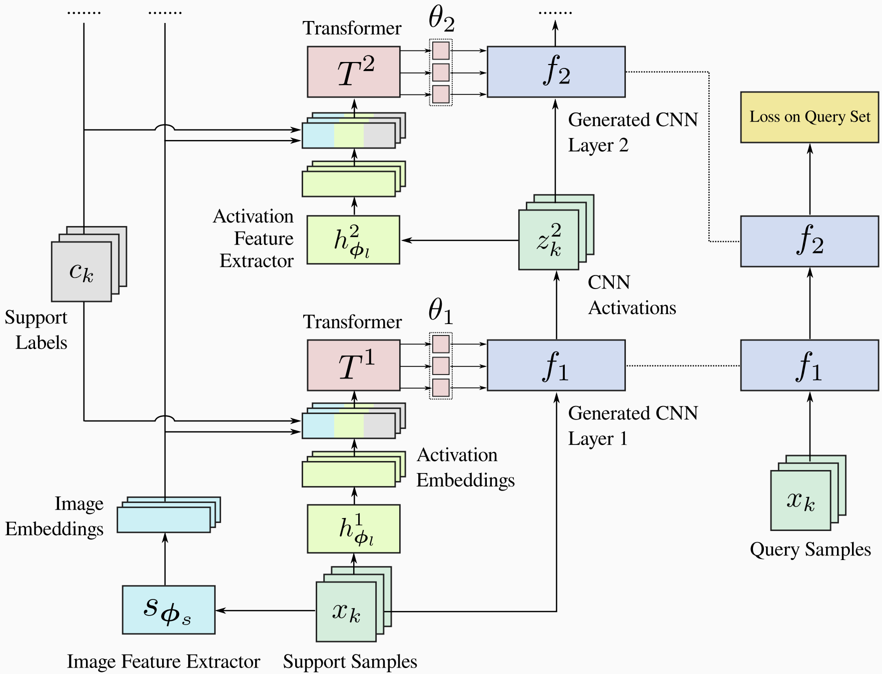
Model Architecture

HyperTransformer Architecture: Generating Each Layer

- Layers are generated **sequentially**
- **For each layer**: sample embeddings → Transformer → weight slices
- **Sample embeddings**: image and activation embeddings
- **Weight embeddings**: "positional encodings" – which slice to be generated at this token



HyperTransformer Architecture: Complete Model



Experimental Results

Observations

- Trained HT **generalizes** to unseen classes
- On small models, HT outperforms RFS, MAML++
  - Dramatically better training accuracy, better test accuracy
- On large models, HT matches performance of SOTA approaches

Dataset	*Approach*	*1-shot (channels)* 8	16	32	48	64	*5-shot (channels)* 8	16	32	48	64
OMNIGLOT	MAML++	81.4	88.6	**95.6**	**95.8**	**97.7**[†]	83.2	94.9	**98.6**	**98.8**	**99.3**[†]
	HT	**87.2**	**93.7**	**95.5**	**95.7**	96.2	**94.7**	**98.0**	**98.6**	**98.8**	98.8
MINIIMAGENET	MAML++	43.9	46.6	49.4	52.2[†]	–	**59.0**	**64.6**	**66.8**	68.3[†]	–
	RFS	44.0	49.4	51.5	54.2	–	56.1	63.5	**67.1**	**69.1**	–
	HT	**45.5**	**50.2**	**53.8**	**55.1**	–	**59.3**	64.2	**67.1**	68.1	–
TIEREDIMAGENET	RFS	44.1	47.7	51.5	54.6	**56.8**[⋆]	55.5	62.0	66.3	69.3	73.2[⋆]
	HT	**49.1**	**51.9**	**54.0**	**55.0**	56.3	**61.9**	**65.8**	**70.2**	**71.1**	**73.9**

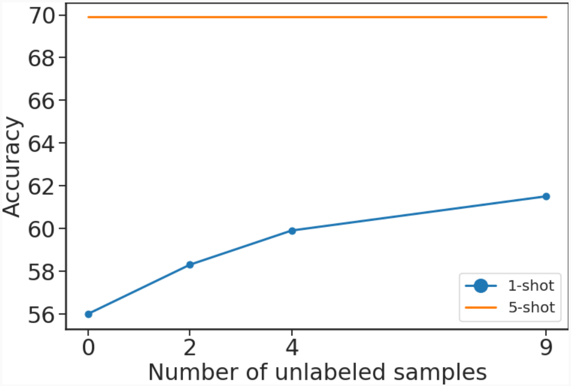
Observations

- Trained HT **generalizes** to unseen classes
- On small models, HT outperforms RFS, MAML++
  - Dramatically better training accuracy, better test accuracy
- On large models, HT matches performance of SOTA approaches

miniImageNet						tieredImageNet		
Method	**1-S**	**5-S**	**Method**	**1-S**	**5-S**	**Method**	**1-S**	**5-S**
HT	54.1	68.5	HT-48	**55.1**	68.1	HT-32	**54.0**	**70.2**
MN	43.6	55.3	SAML	52.2	66.5	MAML-32	51.7	**70.3**
IMP	49.2	64.7	GCR	53.2	**72.3**	HT	**56.3**	**73.9**
PN	49.4	68.2	KTN	54.6	71.2	PN	53.3	72.7
MELR	**55.4**	**72.3**	PARN	**55.2**	71.6	MELR	**56.4**	**73.2**
TAML	51.8	66.1	PPA	54.5	67.9	RN	54.5	71.3

Semi-Supervised Learning: Results

- HT can use **complex task descriptions** as its input
- For example, we used **additional unlabeled samples** to better describe a few-shot learning task
- **2-layer Transformers** were necessary for HT to leverage unlabeled samples
- HT learned to use information about unlabeled samples (*tieredImageNet* 5-way)



Thank You!